Labor_10

SPAM EMAIL CLASSIFICATION

processEmail.m

- Lower-casing

- Stripping HTML

- Normalizing URLs

- Normalizing Email Addresses

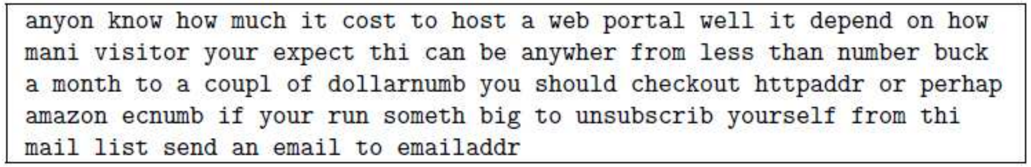
- Nurmalizing Numbers

- Normalizing Dollars

- Word Steamming („include”, „includes”, „included” -> „includ”)

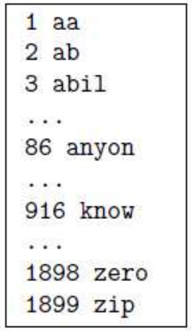
- Removal of non-words

preprocessEmail.m



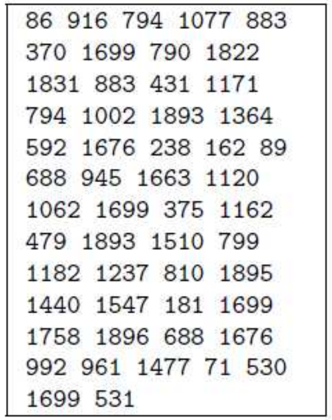
```
anyon know how much it cost to host a web portal well it depend on how
mani visitor your expect thi can be anywher from less than number buck
a month to a coupl of dollarnumb you should checkout httpaddr or perhap
amazon ecnumb if your run someth big to unsubscrib yourself from thi
mail list send an email to emailaddr
```

Figure 9: Preprocessed Sample Email



```
1 aa
2 ab
3 abil
...
86 anyon
...
916 know
...
1898 zero
1899 zip
```

Figure 10: Vocabulary List



```
86 916 794 1077 883
370 1699 790 1822
1831 883 431 1171
794 1002 1893 1364
592 1676 238 162 89
688 945 1663 1120
1062 1699 375 1162
479 1893 1510 799
1182 1237 810 1895
1440 1547 181 1699
1758 1896 688 1676
992 961 1477 71 530
1699 531
```

Figure 11: Word Indices for Sample Email

Vocabulary List

The next step is to choose which words we would like to use in our classifier and which we would want to leave out.

Convert each email into a vector. Specifically whether the i-th word in the dictionary occurs in the email.:

$$x = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n$$